Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Date:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Period:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Using Bioinformatics to Investigate Evolutionary Relationships**  
Adapted from AP Curriculum BLAST Lab

Thank you to Ms. Glick!

**Introduction:** Between 1990-2003, scientists working on an international research project known as the Human Genome Project, were able to identify and map the 20,000 – 25,000 genes that define a human being. The project also successfully mapped the genomes of other species, including the fruit fly, mouse and Escherichia coli. The location and complete sequence of the genes in each of these species are available for anyone in the world to access via the Internet.

Why is this information important? Being able to identify the precise location and sequence of human genes will allow us to better understand genetic diseases. In addition, learning about the sequence of genes in other species helps us understand evolutionary relationships among organisms. Many of your genes are identical of similar to those found in other species.

Suppose you identify a single gene that is responsible for a particular disease in fruit flies. Is that same gene found in humans? Does it cause a similar disease? It would take you nearly 10 years to read through the entire human genome to try to locate the same sequence of bases as that in fruit flies. This definitely isn’t practical, so a sophisticated technological method is needed.

Bioinformatics is a field that combines statistics, mathematical modeling, and computer science to analyze biological data. Using bioinformatics methods, entire genomes can be quickly compared in order to detect genetic similarities and differences. An extremely powerful bioinformatics tool is BLAST, which stands for Basic Local Alignment Search Tool. Using BLAST, you can input a gene sequence of interest and search entire genomic libraries for identical or similar sequences in a matter of seconds.

In this investigation, you will use BLAST to compare several genes, and then use the information to construct a cladogram. A cladogram (also called a phylogenetic tree) is a visualization of the evolutionary relatedness of species. Figure 1 below is a simple cladogram

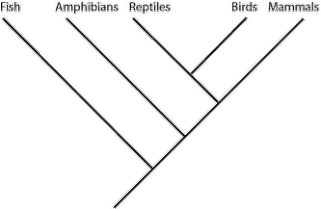
Note that the cladogram is treelike, with the endpoints of each branch representing a specific species. The closer the two species are located to each other, the more recently they share a common ancestor.

Figure 1: A simple cladogram of the Phylum Chordata

The second cladogram in Figure 2 below includes additional details, such as the evolution on particular physical structures called synapomorphies (shared derived characters/traits). Note that the placement of the derived traits corresponds to when (in general, not a specific, sense) that character evolved; every species above the character label possesses that structure. For example, mice and chimps have hair but salamanders and perch do not.

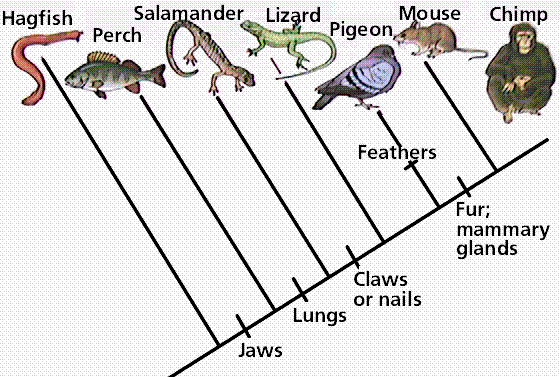


Figure 2: A simple cladogram showing synapomorphies - shared derived traits.

Historically, only physical structures were used to create cladograms; however modern-day cladistics relies heavily on genetic evidence as well. Chimpanzees and humans share over 95% of their DNA, which would place them closely together on a cladogram. Humans and fruit flies share approximately 60% of their DNA, which would place them farther apart on a cladogram.

**Example 1:**

Use the following data to construct a cladogram for the major plant groups below the table. Cladogram can be placed to the right of the chart

Table 1: Physical Characteristics of Major Plant Groups

|  |  |  |  |
| --- | --- | --- | --- |
| Plant Group | Vascular Tissue | Flowers | Seeds |
| mosses | 0 | 0 | 0 |
| pine trees | + | 0 | + |
| flowering plants | + | + | + |
| ferns | + | 0 | 0 |

**Example 2:**

GAPDH (glyceraldehydes 3-phosphate dehydrogenase) is an enzyme that catalyzes the sixth step in glycolysis, an important reaction that produces molecules used in cellular respiration. The following data table shows the percentage similarity of this gene and the protein it expresses in humans versus other species. For example, according to the table, the GAPDH gene in chimpanzees is 99.6% identical to the gene found in humans, while the primary sequence of the corresponding protein is identical.

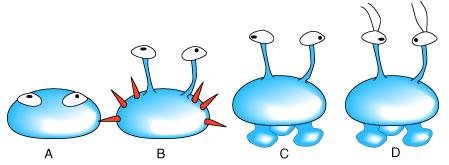
Table 2: Percentage Similarity of the GAPDH Gene and Protein with Homo Sapiens

|  |  |  |
| --- | --- | --- |
| Species | Gene Percentage Similarity with Homo Sapiens | Protein Percentage Similarity with Homo Sapiens |
| Chimpanzee (Pan troglodytes) | 99.6% | 100% |
| Dog (Canis lupis familiaris) | 91.3% | 95.2% |
| Fruit fly (Drosophila melanogaster) | 72.4% | 76.7% |
| Roundworm (Caenorhabditis elegans) | 68.2% | 74.3% |

Questions:

1. Why is the percentage similarity in the gene always lower than the percentage similarity in the protein for each of the species? (Hint: Recall how a gene is expressed to produce a protein. I’m not going to grade this one, but see what you can remember from freshman biology!)
2. Draw a cladogram depicting the evolutionary relationships among all five species (including Humans) according to their percentage similarity in the GAPDH gene.

**Example 3:**

The following four alien species were discovered, and it was realized that species “A” is very primitive – therefore it is the outgroup for the bunch. Use the pictures of each species to answer the following questions.

a. Fill out the following character matrix with a “O” for the ancestral trait (outgroup), and a “+” for a derived trait (different than outgroup).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Physical Characteristic | SPECIES | | | |
| A | B | C | D |
| eyes present |  |  |  |  |
| spines present |  |  |  |  |
| eyes on stalks |  |  |  |  |
| “feet” present |  |  |  |  |
| antennae present |  |  |  |  |

b. Generate a cladogram from the above matrix in the space below. Label the position of the synapomorphies.

**Using BLAST**

A team of scientists has uncovered the fossil specimen in the photo to the left near Liaoning Province, China. You should make some preliminary observations about this fossil based on its morphological features.

Little is known about the fossil. It appears to be a new species. Upon careful examination, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization; however, rare situations of such preservation do occur. Scientists were able to extract DNA nucleotides from the tissue and use the information to sequence several genes.

You will use bioinformatic methods to support a hypothesis as to the evolutionary relationship between modern organisms and the fossilized organism using DNA collected from soft tissues. You will use BLAST to analyze the information from several genes and determine the most likely placement of the fossil species on the figure on the right.

Procedure:

1. Form an initial hypothesis as to where you believe the fossil specimen should be place on the cladogram **Draw your hypothesis on the cladogram below** (Figure 4). Explain your hypothesis based on your observations of the fossil’s morphological characteristics.

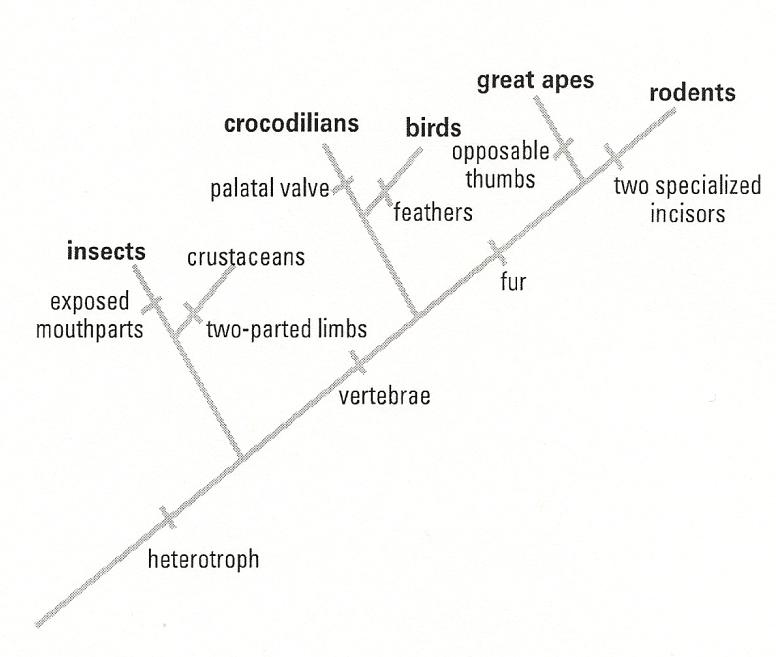


Figure 4: Cladogram of some related groups to the newly discovered fossil found in China.

1. Locate and download gene files 1, 3, and 4 obtained by sequencing the DNA collected from the soft tissue in the fossil. Download 3 files from: http://apcentral.collegeboard.com/apc/members/courses/teachers\_corner/218954.html (When you download these files, click on the folder, and you will see an “asn” file. Drag these “asn” files to the computer Desktop, and you should be able to tell based on the file name, which one is gene 1, 3, and 4. You will not be able to view them but you will be able to upload them into the BLAST site.)
2. Upload each gene sequence (1, 3, and 4) into BLAST by following the directions below. You will be doing this for each sequence separately.
   1. Go to the BLAST site: http://blast.ncbi.nlm.nih.gov/Blast.cgi
   2. Click on “Saved Strategies” from the menu at the top of the page. This will allow you to upload your first file.
   3. Under “Upload Search Strategy,” click on “Choose File” and locate your first file in your file directory.
   4. Click on “View.”
   5. A screen will appear with the parameters for your query already configured. NOTE: Look but do not alter any of the parameters. Scroll down the page and click on the “BLAST” button at the bottom of the page.
   6. Collect and analyze the information from your first gene sequence (according to the instructions on the next page). Then do this with the other gene sequences.

*Analysis of Sequences – Graphic Summary*

The chart below (Figure 5) is a graphical summary of your first sequence. The first line represents the most similar sequence that BLAST was able to pull from the data bases it searched. If you hover over the line, you will see the species of organism that this sequence was derived from. Other, less similar sequences are included from top (most similar) to bottom (less similar).

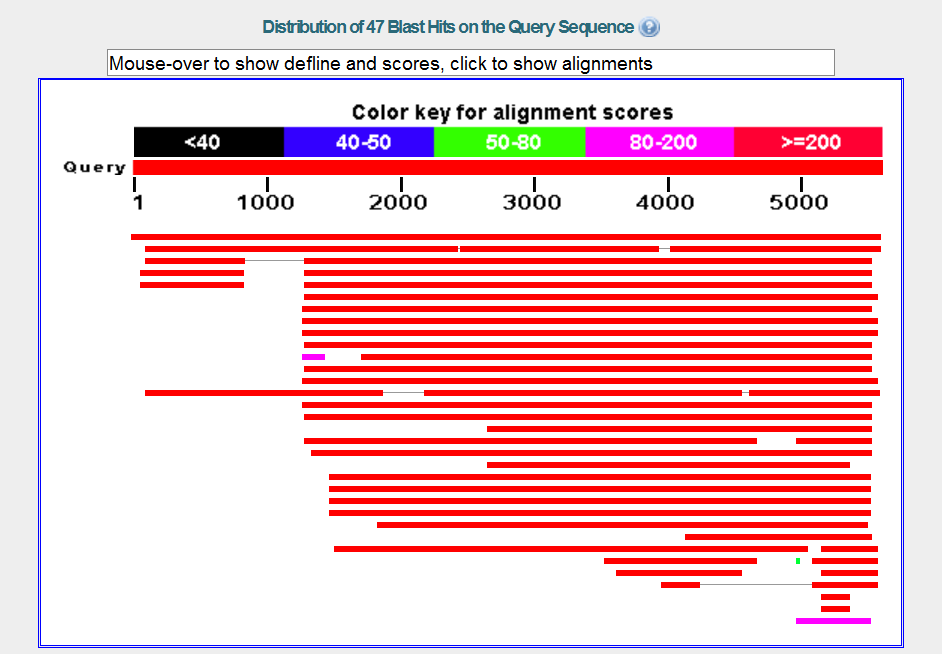


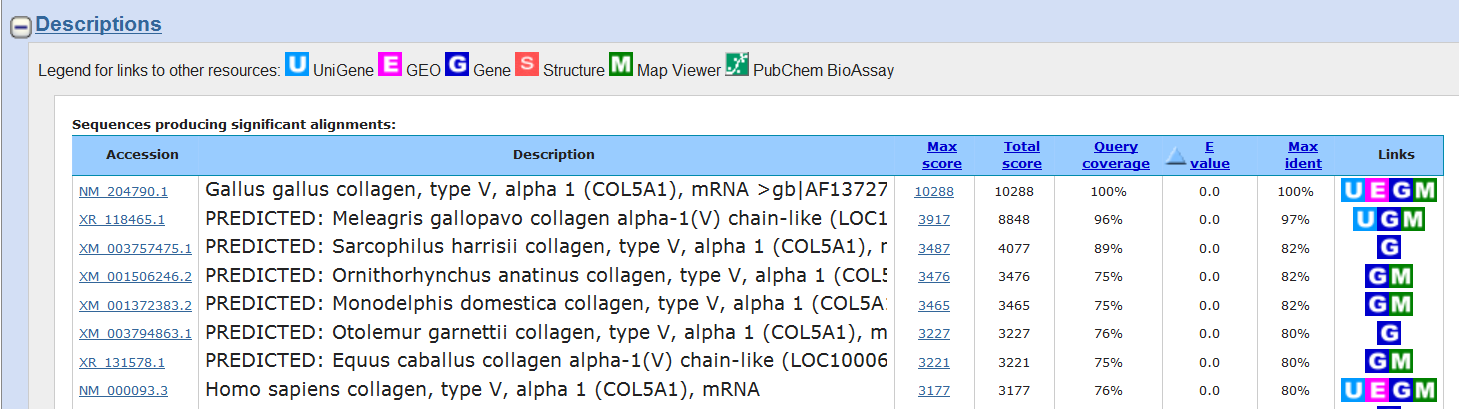
Figure 5: Graphical summary from BLAST.

*Analysis of Sequences – Descriptive Summary*

Table 3 below lists from top to bottom the information about each of the organisms that were represented in the graphic summary. The species in the list are those with sequences identical to or most similar to the gene of interest. The most similar sequences are first, and as you move down the list, the sequences become less similar to your gene of interest.

\*NOTE: Species with common ancestry will share similar gene sequences. The more similar two species’ gene sequences are to one another, the more recent their common ancestor and the closer the two species will be located on the cladogram.

Table 3: Descriptive Summary from BLAST

*  Max(imum) Score: the highest alignment score of a set of aligned segments from the same subject (database) sequence. This normally gives the same sorting order as the E Value. The higher the max score, the closer the alignment (i.e. similarity) between the fossil sequence and the species that is listed. This means the fossil species and the listed species are very closely related to one another. Record the max score of the species with the most similar gene sequence to that of the fossil species in the chart for Analysis Question #1 at the end of this packet.
* E(xpect) Value: The lower the e value, the closer the alignment between the fossil sequence and the species that is listed.
* Accession: If you click on the Accession number for a particular species listed, you will get a full report (Figure 6) that include the classification scheme of the species, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest. It will identify the gene, in this case it is collagen, that you are working with and you will also see the **common name** (ex: domestic dog) of the organism if it has one! Record the scientific name (ex: *Homo sapiens)* AND the common name of the species with the most similar gene sequence to that of the fossil species in the chart for Analysis Question #1 at the end of this packet. Some species’ scientific names will have three parts (ex: *Aquila chrysaetos canadensis*) for the genus identifier (ex: *Aquila*), species identifier (ex: *chrysaetos*), and sub-species identifier (ex: *canadensis)*

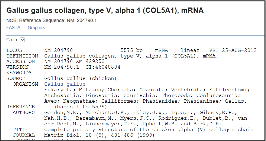


Figure 6: Summary report of a gene from BLAST.

1. Scroll back up to the top of the analysis page. Right above the graphics section, find “Other reports”. Click on “Distance tree of results” to see how this gene aligns with other species. You should see a phylogenetic tree like the one shown below in Figure 7.

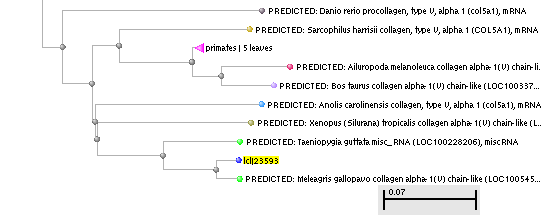


Figure 7: Cladogram generated from BLAST

**Analysis Questions:**

1. Once you have loaded each of your sequences into BLAST, and you have your results, indicate which species in the results has the most similar gene sequence to each of the genes (1, 3, and 4) from the fossil species in the table below.

|  |  |  |
| --- | --- | --- |
| **Gene #** | **Species with most similar gene** | **How similar is it (Max Score)** |
| Gene 1 |  |  |
| Gene 3 |  |  |
| Gene 4 |  |  |

1. Did the analysis of each gene support or refute your original hypothesis? With this new information, where would you place the fossil species on the cladogram on page 5?
2. On the main page of BLAST, click on the link “List All Genomic Databases.” How many genomes (i.e. sequences of all the genes in a particular species) are currently available for making comparisons using BLAST? How does this limitation impact the proper analysis of the gene data used in this lab?